Reprinted from Trends in Biochemical Sciences – *August 1981* **6(8)**, 217–219

Interpretation of amino acid compositions

Athel Cornish-Bowden

Comparing the amino acid compositions of two proteins can give a reliable indication of whether the corresponding sequences are related. When a relationship does exist its extent can be estimated from the compositions and used in the construction of phylogenetic trees.

Leonid Brezhnev[1] has remarked that 'there is nothing more practical than a good theory'. It seems unlikely that he had amino acid compositions of proteins in mind when he said this, but they nonetheless illustrate the usefulness of a good theoretical basis for drawing practical conclusions from experiments. In almost every issue of any general journal of biochemistry there are tables of amino acid compositions, but it is often unclear why they have been measured, as no conclusions, or only the most trivial conclusions, are drawn from them. I shall try to show in this article that a simple theory allows us to deduce whether two protein compositions are similar enough for us to be confident that the sequences are related. Furthermore, when a group of proteins are known to be related to one another one can use their compositions to obtain a phylogenetic tree that compares fairly well with that obtained from the sequences.

Compositions of similar proteins are often found to be similar. Whether the similarities have any statistical significance is, however, usually unclear although several indexes have been proposed during the past twelve years to quantify the amount of difference between compositions[2–6]. The oldest of these is the 'difference index' (DI) of Metzger and co-workers[2], which can be defined as follows:

$$DI = 50\sum\left|\frac{n_{iA}}{N_A} - \frac{n_{iB}}{N_B}\right|$$

where n_{iA} is the number of residues of the ith type in protein A, n_{iB} is the corresponding number in protein B, and N_A and N_B are the total numbers of residues in A and B respectively. The summation is carried out over as many kinds of residue as are distinguished by the measurements (commonly 18, as aspartate normally has to be combined with asparagine and glutamate with glutamine). Three other indexes[3–6] share the following characteristics with the difference index: (1) they can be applied without evident restriction to proteins of quite different molecular sizes; (2) they are defined in terms of mole fractions rather than numbers of residues; (3) they are derived from no theory that might allow their significance to be assessed. It is not surprising, therefore, that these indexes have commonly been reported so tentatively that it has been clear that experimenters have placed little confidence in any predictions that they might make.

Athel Cornish-Bowden is at the Department of Biochemistry, University of Birmingham, P.O. Box 363, Birmingham B15 2TT, U.K.

In the belief that even a very naive theory makes a better starting point than no theory at all, I have proposed a new index $S\Delta n$ that is predicted to behave in a simple way in some circumstances[7]. The symbol is chosen to suggest that $S\Delta n$ is closely related to $S\Delta Q$, the index of Marchalonis and Weltman[4], when that is restricted to comparisons between proteins with the same number of residues. It is defined in terms of the symbols given above as

$$S\Delta n = \tfrac{1}{2}\sum (n_{iA} - n_{iB})^2$$

If both proteins have the same number of residues N, and if one can assume that they have the same statistical properties, then $S\Delta n$ provides an unbiased estimate of the number of differences between the sequences of A and B, with a coefficient of variation of around 38%. The assumption of the same statistical properties is a strong one, because it means not only that one can treat proteins as if they were random sequences of amino acids, but also that the probability of finding any kind of residue at any site is the same for both proteins. This is especially implausible when the proteins are unrelated. Nonetheless, the properties of $S\Delta n$ when applied to proteins of known sequence prove surprisingly close to those predicted[8].

In general, $S\Delta n$ tends to exaggerate the amount of sequence difference, slightly if the proteins are related, but often grossly if they are not. When applied to ten α-chains of haemoglobin, for example, the 45 values of $S\Delta n$ for all possible pairs were on average about 10% higher than the corresponding numbers of sequence differences, but this bias was small compared with the coefficient of variation, which was about 42%, close to the expected value of 38%. Other sets of related proteins (such as insulins, ribonucleases, snake-venom toxins, cytochromes *c* etc.) gave similar results. With unrelated proteins, which typically differ at about 93% of sites, $S\Delta n$ usually exceeds the number of sites; i.e. if interpreted literally it would indicate more than 100% difference. Although these over-estimates are sometimes absurdly large – e.g., human calcitonin and tuna protamine are predicted to differ at about 750% of sites, whereas in reality they differ at only 33 out of 34 – they are unlikely to cause any wrong conclusions to be drawn. Thus, although $S\Delta n$ may fail to indicate a genuine relationship it virtually never indicates a spurious one.

The exaggeration of the amount of difference between unrelated proteins arises from the unrealistic assumption that both proteins have the same statistical properties. If this assumption is not made the same theoretical approach predicts that $S\Delta n$ should be larger than the actual number of sequence differences[8]. Consequently, although under the simplest assumptions $S\Delta n$ values less than $0.42N$ are significant at the 95% level of confidence[7], in practice far fewer than 5% of unrelated pairs given $S\Delta n$ values less than $0.42N$, and $0.93N$ proves to be a more realistic test value[9]. By use of both test values any pair of proteins can be classified with near certainty into one of three categories:

(1) *Related*, if $S\Delta n < 0.42N$;

(2) *Inconclusive*, if $0.42N < S\Delta n < 0.93N$;

(3) *Probably unrelated*, if $S\Delta n > 0.93N$.

The usefulness of this classification would be severely curtailed if many pairs of proteins fell into the inconclusive category. Fortunately, few do[9], at least in the case of unrelated proteins; one cannot make a similar generalization about related proteins because for them the likely range of $S\Delta n/N$ values depends, not surprisingly, on how closely they are related. I have only encountered one pair (out of several thousand examined), namely bovine neurophysin and bovine posterior pituitary peptide, that seem virtually certain to be related on the basis of their compositions yet in fact show no sequence similarity[9].

Despite the good agreement between theory and observation, a plausible objection to the use of $S\Delta n$ for estimating the amount of difference between related proteins is that the statistical uncertainty is so large that the estimate is almost worthless. This would be a valid objection if there

0376 – 5067/81/0000 – 0000/$02.50

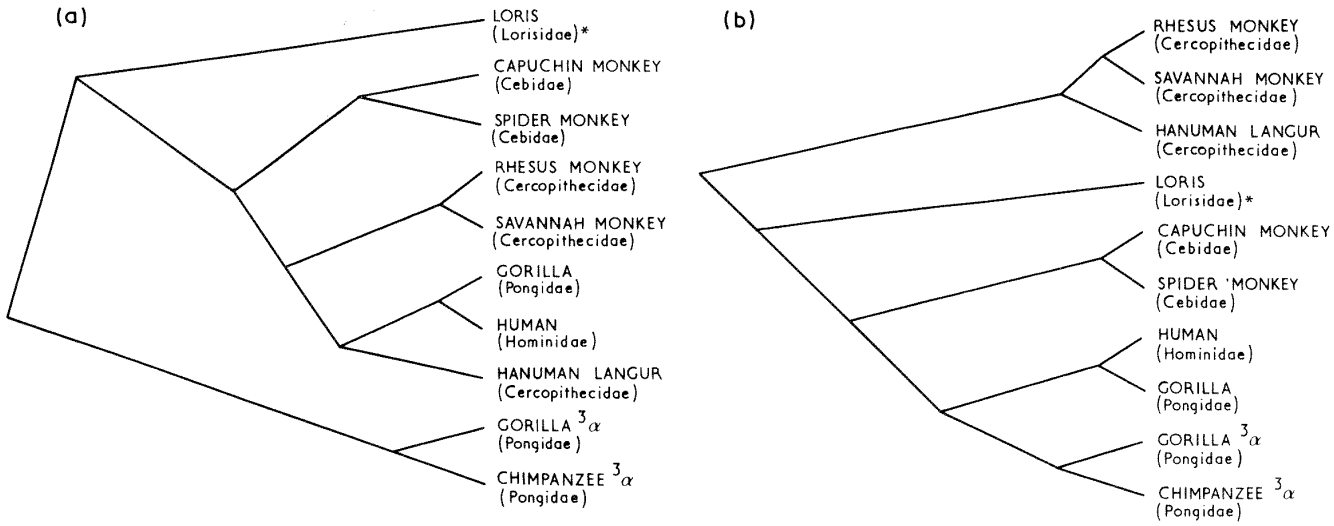

Fig. 1. Comparison of phylogenetic trees obtained from (a) sequences and (b) compositions of the α-chains of haemoglobin. The trees were constructed from data compiled by Hunt and Dayhoff[13]. The $^{3}\alpha$-chains are minor variants observed in chimpanzee and gorilla. Families are shown in parentheses. All are in sub-order Anthropoidiae, *apart from the family* Lorisidae *(*), which is in sub-order* Prosimii. The figure is reproduced with permission of the Biochemical Society from Ref. 12.

were any interest in accurately quantifying the amount of difference between unrelated proteins. But, when the proportion of non-identical sites is around 93% no one really cares whether the true proportion is 88% or 98%: in other words, when $S\Delta n$ gives a very poor numerical estimate of the amount of difference it is of little interest. What one is likely to want to know in practice is the amount of similarity between closely related proteins, and in such cases an imprecise estimate of the amount of *difference* corresponds to a fairly precise estimate of the amount of *similarity*: for example, in 105 comparisons between pairs of myoglobin chains of known sequence[8], the two worst results were both (by co-incidence) obtained with proteins that differed at 19 sites (12% of 153), but one $S\Delta n$ value was 5 whereas the other was 40. However, this eightfold range corresponds to a range of 74 – 97% in the estimated proportion of identical sites. These can hardly be regarded as worthless estimates of a true value of 88%, especially when one remembers that they are not typical results but the worst out of 105.

A major use of protein sequences has been the construction of phylogenetic trees. The globins, for example, have been used extensively in this way[10]. However, given that composition measurements provide very imprecise estimates of sequence differences, one might expect that they would be of little value for phylogenetic purposes. Nonetheless, Black and Harkins[11] have shown that a reasonable tree can be constructed from pyruvate kinase compositions, and in Fig. 1 the trees predicted by the compositions and sequences of the α-chains of ten primate haemoglobins are compared[12]. In the latter example, the sequence tree is generally better than the composition tree, as one would expect, but it is not better in all respects. On the one hand, the composition tree incorrectly places the loris, a member of sub-order *Prosimii*, between the families *Cercopithecidae* (old world monkeys) and *Cebidae* (new world monkeys), both of which are in sub-order *Anthropoidiae*; on the other hand, it classifies the three members of the family *Cercopithecidae* together, whereas the sequence tree wrongly separates them. Moreover, the inclusion of two variant $^{3}\alpha$-chains from gorilla and chimpanzee reminds us that closely related species may contain distantly related proteins: if this is not realized one may get absurd results from protein comparisons, regardless of whether these are based on sequences or on some low-grade alternative, such as compositions, immunological properties or electrophoretic mobility.

Leaving aside the question of distantly related proteins in closely related species, how is it possible for sequences to produce worse results than compositions? The answer almost certainly lies in the role of chance in evolution and in the construction of phylogenetic trees. Chance is especially important when one uses a small number of random substitutions during evolution as a measure of the time since the separation of two lines of descent. For the data used in Fig. 1, only 1–6 sequence differences separate any pair out of gorilla (normal α), human and the three old world monkeys: even if one assumes a constant average rate of substitution, these numbers can at best give estimates of the times of separation that are subject to statistical uncertainties of the order of 40–100%, compared with which the extra uncertainty of 38% associated with $S\Delta n$ seems tolerable. If a single threonine residue in the langur sequence had been a serine residue the results from the sequence classification would have been quite different and much closer to conventional ideas about the relationships between the five species mentioned. It is striking that basic statistical considerations that are commonplace in measurements of radioactivity are apt to be overlooked in discussions of protein evolution. No one, for example, would think that a sample giving a total of three counts was significantly more radioactive than one showing only two, but construction of phylogenetic trees from protein sequences often requires distinction between numbers that are not significantly different.

These considerations lead one to ask how different a group of proteins need to be from one another before the extra labour required for sequence measurements is repaid in better phylogenetic results than those that one could expect to get from the corresponding compositions? This question cannot be answered by studying results with proteins of known sequence, because there are not enough of them, but studies with sequences obtained by simulating the course of evolution provide some indication[12]. They suggest that about 10–15 substitutions need to separate the most distantly related proteins in a group before one can expect sequences to

yield substantially better results than compositions. We can put this into perspective by considering a particular group of species, such as the hoofed mammals, represented by donkey, horse, pig, llama, cow, goat and sheep, and their classification by means of proteins such as histone H2A, cytochrome *c* and haemoglobin. Histone H2A, a very slowly evolving protein, would be expected to give results barely better than classification at random, with no perceptible difference between those from sequences and those from compositions. Cytochrome *c* would give appreciably better than random results, but would show little difference between sequences and compositions. Only with a rapidly evolving protein such as haemoglobin could one expect nearly perfect results with sequences and appreciably worse ones with compositions.

Since its inception *TIBS* has devoted much space to the development of new techniques in biochemistry. This is as it should be, but occasionally it is also useful to examine well established techniques and to enquire as to whether they are providing as much information as they might. Amino acid compositions provide an example of a kind of measurement that is capable of providing useful information but is rarely used for anything apart from filling space in journals.

References

1 Brezhnev, L., quoted by V. Rich (1977) in *Nature (London)* 270, 470–471

2 Metzger, H., Shapiro, M. B., Mosimann, J. E. and Vinton, J. E. (1968) *Nature (London)* 219, 1166–1168

3 Harris, C. E., Kobes, R. D., Teller, D. C. and Rutter, W. J. (1969) *Biochemistry* 8, 2442–2454

4 Marchalonis, J. J. and Weltman, J. K. (1971) *Comp. Biochem. Physiol.* 38B, 609–625

5 Harris, C. E. and Teller, D. C.(1973) *J. Theor. Biol.* 38, 347–362

6 Dedman, J. R., Gracy, R. W. and Harris, B. G. (1974) *Comp. Biochem. Physiol.* 49B, 715–731

7 Cornish-Bowden, A. (1977) *J. Theor. Biol.* 65, 735–742

8 Cornish-Bowden, A. (1979) *J. Theor. Biol.* 76, 369–386

9 Cornish-Bowden, A. (1980) *Anal. Biochem.* 105, 233–238

10 Goodman, M. (1976) in *Molecular Evolution* (Ayala, F. J., ed.), pp. 141–159, Sinauer Associates, Sunderland, Massachusetts

11 Black, J. A. and Harkins, R. N. (1977) *J. Theor. Biol.* 66, 281–295

12 Cornish-Bowden, A. (1980) *Biochem. J.* 191, 349–354

13 Hunt, L. T. and Dayhoff, M. O. (1976) in *Atlas of Protein Sequence and Structure* (Dayhoff, M. O., ed.), vol. 5, suppl. 2, pp. 191–223, National Biomedical Research Foundation, Silver Spring, Maryland